DOCUMENT RESUME

ED 438 329 TM 030 646

AUTHOR Smith, A. Delany; Henson, Robin K.
TITLE State of the Art in Statistical Significance Testing: A Review of the APA Task Force on Statistical Inference's Report and Current Trends.
PUB DATE 2000-01-28
NOTE 49p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Dallas, TX, January 27-29, 2000).
PUB TYPE Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Hypothesis Testing; Research Methodology; Research Reports; *Social Science Research; Statistical Inference; *Statistical Significance
IDENTIFIERS American Psychological Association

ABSTRACT

This paper addresses the state of the art regarding the use of statistical significance tests (SSTs). How social science research will be conducted in the future is impacted directly by current debates regarding hypothesis testing. This paper: (1) briefly explicates the current debate on hypothesis testing; (2) reviews the newly published report of the American Psychological Association Task Force on Statistical Inference; (3) examines current trends in reporting practices in journals; and (4) presents recommendations for researchers to advance scientific inquiry in the social sciences. (Contains 64 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Running Head: STATE OF THE ART

ED 438 329

State of the Art in Statistical Significance Testing: A Review of the APA Task Force on Statistical Inference's Report and Current Trends

A. Delany Smith and Robin K. Henson

The University of Southern Mississippi

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Robin K. Henson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

BEST COPY AVAILABLE

TM030646

Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, January 28, 2000.
Correspondence concerning this manuscript should be sent to the second author at Department of Educational Leadership and Research, Box 5027, Hattiesburg, MS 39401-5027; or via email at robin.henson@usm.edu.

Abstract

The present paper addresses the state of the art regarding the use of statistical significance tests (SSTs). How social science research will be conducted in the future is directly impacted by current debates regarding hypothesis testing. This paper (a) briefly explicates the current debate regarding hypothesis testing, (b) reviews the newly published APA Task Force on Statistical Inference's report, (c) examines current trends concerning reporting practices in journals, and (d) presents recommendations for researchers to advance scientific inquiry in the social sciences.

State of the Art in Statistical Significance Testing: A Review of the APA Task Force on Statistical Inference's Report and Current Trends

Statistical significance testing (sometimes called null hypothesis testing) has historically dominated research statistical methods in social science research (cf. Daniel, 1998; Huberty, 1993; McLean & Ernest, 1998; Nix & Barnette, 1998). More recently, however, the utility of significance testing as means of testing research effects has been severely questioned. The debate has been rather furious. For example, Harris (1991) noted:

> There has been a long and honorable tradition of blistering attacks on the role of statistical significance testing in the behavioral sciences, a tradition reminiscent of knights in shining armor bravely marching off, one by one, to slay a rather large and stubborn dragon. . . . Given the cogency, vehemence and repetition of such attacks, it is surprising to see that the dragon will not stay dead. (p. 375)

The "dragon" still lives for multiple reasons. Some researchers have come to the defense of statistical significance

testing (cf. Abelson, 1997; Cortina & Dunlap, 1997; Levin, 1998a, 1998b) and argued for its rightful place in the social scientist's arsenal. Vacha-Haase and Thompson (1998) have suggested that statistical significance testing continues to be used largely because of the steeped tradition it holds; it is what most researchers know. They also argued that researchers will not change their ways until journal editors require them to do so, suggesting that old habits die hard and perhaps not without some extrinsic motivation.

Indeed, Pedhazur and Schmelkin (1991) noted that "Probably few methodological issues have generated as much controversy among sociobehavioral scientists as the use of [statistical significance] tests" (p. 198). Pedhazur (1997) indicated that the "controversy is due, in part, to various misconceptions of the role and meaning of such [statistical significance] tests in the context of scientific inquiry" (p. 26). These "misconceptions" have been attacked for considerable time (see e.g., Berkson, 1938; Tyler, 1931), and yet they persist in modern research practice.

In the midst of the fray, the American Psychological Association (APA) convened a committee, the Task Force on Statistical Inference (TFSI), to examine current statistical

practices, including statistical significance testing (Azar, 1997). Recently, the TFSI published its final report in American Psychologist (Wilkinson & TFSI, 1999) to generate discussion concerning the matters therein. This report is intended to affect current practice regarding (among other things) statistical significance testing and, potentially, may impact an official APA position in the next edition of the Publication Manual of the American Psychological Association (American Psychological Association, 1994), due out in 2001.

The purpose of the present paper is to address the state of the art regarding the use of statistical significance tests (referred to as SSTs). How social science research will be conducted in the future is directly impacted by current debates regarding hypothesis testing; and, the state of the art is evolving. This paper (a) explicates the current debate regarding hypothesis testing, (b) reviews the newly published APA TFSI's report on statistical inference, (c) examines current trends concerning reporting practices in journals, and (d) presents recommendations for researchers to advance scientific inquiry in the social sciences.

A Brief Review of the Current Debate

While a comprehensive review of the literature regarding SSTs is beyond the scope of this paper, a brief review is presented here with emphasis on the perspectives of SST opponents. The intent is to highlight some of the key elements of the debate. For more complete reviews of both sides of the issue, the reader is referred to Burdenski (1999); Harlow, Mulaik, and Steiger (1997); and a recent special issue of Research in the Schools (1998, Vol. 5, No. 2).

Low p-value ≠ Replicable Result

Central to the arguments of statistical significance opponents is that the familiar p-value does not suggest the probability of attaining similar results in future samples (Daniel, 1998). In fact, the p-value only indicates the probability (0 to 1.0) of attaining the presently observed results from the present sample assuming that the null hypothesis is exactly true in the population (Thompson, 1994a). As Daniel (1998) noted, "Despite misperceptions to the contrary, the logic of statistical significance testing is NOT an appropriate means for assessing replicability (Carver, 1978; Thompson, 1993a)" (p. 25, emphasis in original). Furthermore, Burdenski (1999, p. 16) suggested, "While psychologists want to

know about the population to determine if the results will generalize and replicate, statistical tests do not provide that information." Instead SSTs only answer the question: Assuming the null hypothesis is true, what is the likelihood of obtaining my results? SSTs do not answer the question: Given my results, what is the likelihood that the null hypothesis is true in the population? As explained by Kirk (1996),

> In scientific inference, what we want to know is the probability that the null hypothesis (H_0) is true given that we have obtained a set of data (D); that is, $p(H_0|D)$. What null hypothesis significance testing tells us is the probability of obtaining these data or more extreme data if the null hypothesis is true, $p(D|H_0)$. Unfortunately for researchers, obtaining data for which $p(D|H_0)$ is low does not imply that $p(H_0|D)$ also is low. (p. 747)

Therefore, the p-value does not indicate the probability of replicability, no matter how many zeros find themselves after the decimal. In reference to researchers' desire to answer the replicability question, Cohen (1994) poignantly noted that a null hypothesis test "does not tell us what we want to know, and we so much want to know what we want to know, out of desperation, we nevertheless believe that it does!" (p. 997).

Sample Size and a False Null

As suggested, in order to assess the likelihood of obtained results it is necessary to assume the null to be true in the population, as a reference point of sorts. The inference is from the population to the sample, not from the sample to the population (as one would hope). A problem with the assumption of a true null is that, in reality, the null seldom (if ever) is exactly true (i.e., there is always at least a minute difference between groups or at least some relationship between variables, however small), and that with enough subjects it will always be rejected (Cohen, 1994; McLean & Ernest, 1998). A p-value, then, speaks not to potential replicability of results, but rather to how large one's sample is given the observed effect. Regarding this issue, Thompson (1998a) noted that "if we fail to reject [the null hypothesis], it's only because we've been too lazy to drag in enough subjects" (p. 799). While Hagen (1997) criticized Cohen's (1994) claim that the null hypothesis will always be found false at some sample size, the point will never be ultimately resolved since the "population is infinite and unknowable" (Burdenski, 1999, p. 17).

However, as Meehl (1978) pointed out, "As I believe is generally recognized by statisticians today and by thoughtful

social scientists, the null hypothesis, taken literally, is always false" (p. 822). Indeed, respected statistician John Tukey (1991) noted, "the effects of A and B are always different - in some decimal place - for any A and B. Thus asking 'Are the effects different?' is foolish" (p. 100). This conclusion is based on the intuitive assumption social science variables are seldom, if ever, completely unrelated to each other, regardless of whether we know why a relationship exists. As Hays (1981) noted, "There is surely nothing on earth that is completely independent of anything else. The strength of association may approach zero, but it should seldom or never be exactly zero" (p. 293). Therefore, Snyder and Thompson (1998) concluded that "nonzero sample effects are always expected" (p. 338). Thirty years ago, Nunnally (1960) observed the connections between the false null, sample size, and statistical significance: "If the null hypothesis is not rejected, it is usually because the N is too small. If enough data are gathered, the hypothesis will generally be rejected" (p. 643).

Practical Versus Statistical Significance

Opponents of SSTs often argue that a statistically significant result (i.e., when the obtained p-value is less than the predetermined alpha level) only indicates the likelihood of

the result and not the importance of a result (cf. Thompson, 1994a). As suggested above, even minute differences between groups will be statistically significant at some sample size. Poor researcher judgment is reflected in "the ingenuous assumption that a statistically significant [or, unlikely] result is necessarily a noteworthy result" (Daniel, 1997, p. 106). Shaver (1985, p. 58) illustrated this point, suggesting that picking up a telephone and being connected to the caller without the telephone ever ringing is certainly unlikely, but also certainly unimportant.

Several issues appear to contribute to this misconception. First, SSTs have historically been used to provide researchers with a dichotomous "reject" or "fail to reject" decision making procedure to evaluate results. In this context, researcher judgment supposedly is set aside in lieu of an "objective" means to determine whether or not a difference or relationship existed in the data. However, Thompson (1999) and McLean and Ernest (1998) explained that social science is ultimately subjective in nature. It is important to remember, for example, that an alpha level is (appropriately) set prior to the study and should be reflective of a risk tolerance for Type I error. Such a judgment, if made thoughtfully, is certainly not "objective" at

all, but rather indicative of careful consideration of study characteristics and the nature of the topic investigated. Unfortunately, such reflective practice has not been historically evident. Cohen (1994) called this errant process ". . . the ritual of null hypothesis significance testing - mechanical dichotomous decisions around a sacred .05 criterion . . ." (p. 997). Researchers' dependence on SSTs to provide "clear cut yes-no decisions" (Cohen, 1990, p. 1307) may lead some to believe that a "yes" decision is necessarily important without consideration of factors that may have contributed to that decision (e.g., sample size, power, effect size, alpha level), or consideration of how large the tested difference or relationship is.

Abelson (1997) defended the use of SSTs to make such categorical statements, noting that they provide a means of entry for knowledge into a field. This knowledge can grow by comparing results across studies or via discussion and reaction to published articles. Frick (1996) also supported such categorical decisions as a criterion for entry into a field of knowledge. Frick suggested this role of SSTs was similar to entry into the baseball hall of fame, where a player must receive 75% of sportswriters' votes for entry. Entry is either

accepted or rejected. However, while use of SSTs certainly do serve a gatekeeping function for recognized research, this function still does not speak to the relative importance of a given finding, but only to the likelihood of obtaining one's statistic (assuming the null is true in the population).

A second reason why a statistically significant result may be considered necessarily important (in and of itself) is semantical. Thompson (1994a) noted: "Many of the problems in contemporary uses of statistical significance testing originate in the *language* researchers use" (p. 6, emphasis in original). Specifically, the use of the term "statistical significance" can be misleading and confused with the general meaning of the term "significant" (i.e., important). While grammatically subtle, it is common to see results referred to as simply "significant", which over time may come to be seen as "important". To foster precision in language use, Daniel (1998) suggested that journal editors "Require authors to use 'statistically' before 'significant'" (p. 29). Some journals have established such policies (see e.g., Thompson, 1994b, Educational and Psychological Measurement). Levin (1993, 1998b), on the other hand, warned that this practice may be little more than policing of language. He noted it is "difficult to support . . .

requirements that take away certain freedoms of author style and expression; in particular, when editorial policy is only half a vowel away from turning into editorial police" (Levin, 1998b, p. 44). While Levin's concerns are merited, the misconceptions surrounding SSTs are equally egregious.

Researcher Misconceptions

At a fundamental level, several empirical studies have indicated that many researchers that rely on SSTs simply do not fully understand what SSTs can and cannot do (cf. Nelson, Rosenthal, Rosnow, 1986; Oakes, 1986; Rosenthal & Gaito, 1963; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993). Even some major statistical textbooks do not present a complete understanding of the limitations that accompany SSTs (Carver, 1978). Some of the misconceptions have been presented above. Interested readers are referred the above citations for a more complete discussion.

Alternative Analyses and Reporting Practices

In light of the perceived weaknesses (perhaps, uselessness according to some) of SSTs, many researchers have argued that other statistical methods should be utilized when evaluating data. Chief among the recommendations is the reporting of some measure of effect (cf. Kirk, 1996; Snyder & Lawson, 1993;

Thompson, 1994b) and the use of confidence intervals (Nix & Barnette, 1998). Schmidt (1996), a vocal critic of SSTs and advocate of alternative reporting methods, insisted that,

> We must abandon the statistical significance test. In our graduate programs we must teach that for analysis of data from individual studies, the appropriate statistics are point estimates of effect sizes and confidence intervals around these point estimates. We must teach that for analysis of data from multiple studies, the appropriate method is meta-analysis. (p. 116)

Others have taken a more moderate stance, emphasizing the need to report effect sizes without calling for the outright ban of SSTs (Daniel, 1998; McLean & Ernest, 1998; Thompson, 1998b). At a minimum, most participants in the debate have agreed that SSTs are abused (misused) in current practice.

The APA Task Force on Statistical Inference's Report

As noted, the American Psychological Association (APA) responded to the debate surrounding SSTs by convening a committee to examine issues related to improved research practice and statistical inference. Recently published, the report (Wilkinson & TFSI, 1999) provides multiple guidelines concerning various research matters. This report, and ensuing

discussion, may impact APA positions on statistical reporting practices. The TFSI's report covers myriad research topics, ranging design considerations to proper use of tables and figures. Imbedded in this ten-page discussion is a brief section concerning hypothesis tests, effect sizes, and interval estimates that bears on the current paper.

Regarding hypothesis tests, the TFSI (Wilkinson & TFSI, 1999) noted, "It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual p value or, better still, a confidence interval. . . . Always provide some effect-size estimate when reporting a p value" (p. 599, emphasis added). Furthermore, the TFSI suggested that researchers "Always present effect sizes for primary outcomes. . . . It helps to add brief comments that place these effect sizes in a practical and theoretical context" (p. 599, emphasis added).

The TFSI report takes a noteworthy step beyond the APA position taken in the fourth edition of the APA publication manual (APA, 1994) where reporting effect sizes is merely "encouraged" (p. 18). The report recognizes that this encouragement has done little to change practice in the field and cites three empirical studies as evidence (Keselman et al.,

1998; Kirk, 1996; Thompson & Snyder, 1998). Importantly, the report also stressed that,

> reporting and interpreting effect sizes in the context of previously reported effects is essential to good research. It enables readers to evaluate the stability of results across samples, designs, and analyses. Reporting effect sizes also informs power analyses and meta-analyses needed in future research. (Wilkinson & TFSI, 1999, p. 599)

The definitive statement to "always" report effect sizes, at least for primary outcomes, is included for the first time in this recent report and represents an important advancement in the field. This practice parallels that of several journal policies already in place. Educational and Psychological Measurement (Thompson, 1994b), Journal of Applied Psychology (Murphy, 1997), Journal of Experimental Education (Heldref Foundation, 1997), and Measurement and Evaluation in Counseling and Development (1992) all have policies that either require or strongly encourage effect size reporting, sometimes in addition to p-values.

While it is clear that the committee's position is in favor of reporting effect sizes, the TFSI stops short of placing an outright ban on statistical significance tests. Of course, this

outcome was foreshadowed by the committee's draft report that noted the TFSI "does not support any action that could be interpreted as banning the use of null hypothesis significance testing" (Board of Scientific Affairs, 1996, p. 1). It appears that the TFSI has taken a moderate approach to the SST issue, and merely "upgraded" the previous "encouragement" to report effect sizes to a more definitive requirement. This conclusion is supported by the overall lack of attention the report gives to a matter that Pedhazur and Schmelkin (1997, p. 26) called "a major source of controversy among social scientists."

What is notably absent from the report is any substantive information that addresses the myriad misconceptions about what SSTs can and cannot do. The misuse and misinterpretation of SSTs has been empirically verified and is a central component of arguments against the use of SSTs. It is possible, perhaps, that the TFSI supposes that the reporting of effect sizes will indirectly bring to light prior misconceptions held by researchers. However, given the steeped tradition of SSTs and their time-honored place in social science research, it is unlikely that indirect approaches will impact practice (cf. Vacha-Haase & Thompson, 1998).

Also absent from the TFSI's report is comment regarding the use of language when referring to a statistically significant result. However, in an extensive section on dealing with multiple outcomes that immediately follows the section on hypothesis tests and effect sizes, the report uses only the word "significant" or "significance" when referring to results, thereby taking an indirect position on language use. Oddly, each time the word "significant" is used, it is always placed in quotes (as above), somehow suggesting that the TFSI is at least aware of the language debate.

It remains to be seen, of course, whether the TFSI's report will impact both the next edition of the APA publication manual and/or the general practice of the field. Given the report's specific reference to "always" report effect sizes, and general professional support for effect size reporting, it is anticipated that the new manual will reflect such change. Indeed, after reviewing the literature, McLean and Ernest (1998) claimed they "were unable to find an article that argued against the value of including some form of effect size or practical significance estimate in a research report" (p. 18).

As for reporting practice, it is unlikely that researchers will enact this change until journal editors require them to do

so (Vacha-Haase & Thompson, 1998). Daniel (1998) hypothesized that, "If improvements are to be made in the interpretation and use of SSTs, professional journals (Rozeboom, 1960), and, more particularly, their editors will no doubt have to assume a leadership role in the effort" (p. 27). APA journals may be the first to follow suit since they are closely tied to the publication manual; other journals may follow behind as the field continues to evolve. Of course, as noted, some journals already have established policies requiring effect size reporting. Shaver (1993) suggested that, "As gatekeepers to the publishing realm, journal editors have tremendous power . . . [and should] become crusaders for an agnostic, if not atheistic, approach to tests of statistical significance" (pp. 310-311).

Finally, the TFSI's report concludes with an acknowledgement of the role of informed researcher judgment versus blind adherence to statistical methods (e.g., using a *p*-value to reject the null without considering the magnitude of effect). Correctly, Wilkinson and the TFSI (1999) stated, "Good theories and intelligent interpretation advance a discipline more than rigid methodological orthodoxy. . . . Statistical methods should guide and discipline our thinking but should not determine it" (p. 604).

Current Trends in Statistical Significance and Effect Size Reporting

Current trends suggest a slow but decisive movement toward reporting effect size measures and using SSTs more accurately. The movement is slow in that authors still seldom include effect size measures and often misuse or misinterpret SSTs. Carver (1993) suggested, "In educational and psychological research, testing for statistical significance has abated very little, if at all, in the 15 years since I presented a lengthy case against these tests" (p. 292). The movement is decisive in that the literature is now replete with arguments against misuse of statistical significance testing (cf. Harlow, Mulaik, & Steiger, 1997) and arguments for proper inclusion of effect size estimates along with other recommended alternative practices, including the APA TSFI report discussed above (cf. Kirk, 1996; Snyder & Lawson, 1993).

It is one thing to claim that current trends reflect a misuse of SSTs and underutilization of effect sizes, and quite another to empirically verify such claims. However, the argument is not mere rhetoric, and indeed, several <u>empirical</u> studies regarding reporting practices for SSTs and effect sizes suggest continued problems in the field. Recent meta-analytic

studies of articles in prominent educational and psychological journals reveal these problems with some clarity. Several of these studies are reviewed here.

Regarding effect size reporting, Kirk (1996) reviewed articles using inferential statistics in the 1995 volumes of four APA journals and found "considerable variability among the journals" (p. 752). The percentage of articles that reported at least one effect size estimate ranged from 12% in the Journal of Experimental Psychology to 77% in the Journal of Applied Psychology. Furthermore, R^2 or some generic reference to a variance-accounted-for statistic accounted for 60% of the effect sizes reported. While the reporting rate for the Journal of Applied Psychology is encouraging, Kirk noted that authors in this journal are more likely to use correlational or regression analyses, which generally yield an R^2 value in common statistical packages. On the other hand, authors in the Journal of Experimental Education tend to utilize ANOVA-type analyses, for which statistics packages less frequently report effect size estimates. As such, one may question whether these reporting practices are a function of what the computer prints out rather than thoughtful researcher judgment. Additionally, Kirk did not indicate how many of the reported effect sizes were actually

interpreted as opposed to being simply listed along with many other statistics some obscure table. A more appropriate measure of effect size use would include an assessment of whether researchers both report and interpret their obtained effect sizes.

Thompson and Snyder (1997) analyzed 22 research articles from two volumes of the Journal of Experimental Education. Almost all of the articles labeled their results as "significant" rather than the more precise "statistically significant." Multiple authors also unfortunately (and incorrectly) described their results as "approaching significance" or being "nearly significant." Only three articles reported evidence of external replicability with independent samples and none reported internal replicability evidence (e.g., via a bootstrap, jackknife, or crossvalidation analysis). This finding suggests either that researchers were (a) unconcerned about replicability evidence, (b) did not know means by which to evaluate replicability, or (c) incorrectly (most likely?) assumed that the familiar p-value served as a measure of replication. As noted above, this last assumption is a common misconception concerning SSTs. Thompson and Snyder also noted that eight articles reported no measures of effect

size, six articles included effect sizes but did not interpret them, and four articles inconsistently presented at least one effect size but did not list estimates of effect for other SSTs in the article. Unlike Kirk (1996), Thompson and Snyder (1997) evaluated both effect size reporting and interpretation. They found that variance-accounted-for effect sizes were used in result presentation and interpretation in only 4 of the 22 articles.

Because large sample sizes allow statistically significant results with small effects and, conversely, large effects are needed for statistically significant results from small sample sizes, Thompson and Snyder (1997) hypothesized that, "(a) disproportionately large effect sizes might occur in studies with smaller sample sizes and that (b) studies with larger sample sizes might involve disproportionately small effect sizes" (p. 79). After calculating effect sizes for all of the studies, Thompson and Snyder reported that the mean effect sizes in the three studies with the largest ns (1,512, 9,987, and 12, 121) "were 14.5%, 28.0%, and 02.2%, respectively" (p. 79). Furthermore, they noted that "6 of the 133 effect sizes from studies with smaller samples (i.e., less than 500) involved

effect sizes analogous to r^2 greater than 60%" (p. 79). In general, Thompson and Snyder (1997) noted that,

> These results are mixed as to whether practice in the journal reflects the movement of the field. The pattern is most favorable with respect to effect size reporting and interpretation, but less favorable with respect to language use and replicability of analyses. (p. 81)

When reviewing 68 research based articles (1990 to 1996) from the primary journal of the Association for Assessment in Counseling of the American Counseling Association, Measurement and Evaluation in Counseling and Development, Vacha-Haase and Nilsson (1998) reported that 81.9% of the articles used SSTs as the basis for result interpretation. Use of the term "statistical significance" was more frequent (33.8%) than in the Thompson and Snyder (1997) study, but authors still only used correct language a third of the time. Interestingly, the term "statistical significance" was used much more frequently in 1990 than 1996. Vacha-Haase and Nilsson also reported that only 35.3% of the articles indexed results to obtained effect sizes and even fewer indexed results to sample sizes (7.3%), which, as noted, dramatically impacts the outcome of SSTs. Furthermore, most articles (86.8%) failed to report the selected alpha level

for the SSTs used - a fundamental error given that alpha is an *a priori* decision regarding risk of Type I error. Vacha-Haase and Nilsson noted that their results "suggest that authors are increasingly using statistical significance testing in their research articles. . . Statistical significance testing continues to be prevalent despite warnings of misunderstandings and misuses (e.g., Carver, 1978; Thompson, 1996)" (p. 54).

Snyder and Thompson (1998) analyzed 35 research articles (Vols. 5 through 11) found in *School Psychology Quarterly*, the official APA Division 16 journal. The authors calculated variance-accounted-for effect sizes (when not reported) for the 321 SSTs used in the articles. The mean effect size was moderate ($M = .13$, $SD = .16$; cf. Cohen, 1988).

Again, precise language use was a problem. Snyder and Thompson (1998) reported that "authors of only five of the 35 articles used the term 'statistically significant' rather than 'significant'" (p. 342). Regarding effect sizes, 19 of the 35 articles reported effect size indices. However, Snyder and Thompson explained that "few authors interpreted these indices" and that the "preponderance of the authors emphasized tests of statistical significance to determine if their results were noteworthy" (p. 342). They also found examples of result

overinterpretation, in which small effects were deemed statistically significant (and thereby noteworthy) due to large sample sizes.

Only two of the articles in Snyder and Thompson's (1998) study conducted an internal replicability analysis, suggesting dependence on the p-value as a measure of whether similar results can be found in future samples (which, of course, it is not). True external replications (with independent samples) were conducted in two articles. Finally, when authors failed to reject their null hypotheses, most authors also failed to "conduct power analyses to determine whether their results were artifacts of small sample size" (Snyder & Thompson, 1998, p. 342).

Keselman et al. (1998) conducted an extensive review of articles found in the 1994 or 1995 issues of 17 prominent journals. Keselman et al. noted, "These journals were chosen because they publish empirical research, are highly regarded within the fields of education and psychology, and represent different education subdisciplines" (p. 353). As such, the authors' review can reasonably be considered comprehensive and reflective of general practice in the field.

Keselman et al. (1998) divided their review by type of analyses reported: between-subjects univariate designs, between-subjects multivariate designs, repeated measures designs, and covariance designs. Unfortunately, effect sizes were seldom reported and power considerations (which directly speak to the relationship between sample size and SSTs) were also infrequent. For between-subjects univariate designs, only 10 of 61 articles considered power or effect sizes, and only 6 of these articles (9.8%) calculated effect sizes directly. For between-subjects multivariate designs, effect size indices were given in 8 of 79 articles (10.1%). Furthermore, most articles reported incomplete SST information, such as not identifying the test criterion (e.g., Wilks) or not including the degrees of freedom with an *F* statistic. Only 16 of 226 articles (7.1%) employing a repeated measures design calculated an effect size, typically Cohen's (1988) *d*. Positively, three articles indicated that non-statistically significant findings found therein may have been due to low power, although none of the articles reported actual assessments of power. Finally, in covariance designs, Keselman et al. reported that 11 of 45 articles (24.4%) included at least one effect size estimate. No articles included results in terms of confidence intervals, a common recommendation to

place SSTs in context (and, hopefully, facilitate appropriate interpretation). Regarding the suggested use of effect sizes and confidence intervals, Keselman et al. noted, "In the present sample of studies, the behavioral science researchers were either unaware of these recommendations or chose to ignore them" (p. 376).

In sum, Keselman et al. (1998) observed:

> As anticipated, effect sizes were almost never reported along with p values, despite encouragement to do so in the most recent edition of the American Psychological Association's (1994) Publication Manual. Moreover, indications of the magnitude of interaction effects were extremely rare. Finally, it should be noted that, in all instances in which effect sizes were given, a statistically significant result was obtained. (p. 358)

Furthermore, the authors "strongly encourage . . . routinely reporting measures of effect" to improve practice and suggested that effect indices are necessary to "distinguish between those results that are 'practically' significant and those that are only 'statistically' significant" (Keselman et al., 1998, pp. 358-359).

Several themes emerged from this review. First, it is clear that SSTs continue to be misrepresented and misused, as reflected in imprecise language use, overinterpretation of small effects with large samples, and dependence on the p-value as a measure of replicability. Second, despite numerous arguments to do so (cf. Cohen, 1994; Kirk, 1996) and encouragement in the APA Publication Manual (APA, 1994), effect sizes are seldom reported and even less frequently interpreted. A total of 24 major journals were examined by the studies reviewed here (Note: Three journals were dually reviewed by two of the studies [Kirk, 1996; Keselman et al., 1998] but it is unclear whether the same issues were examined.) From these journals, a total of 927 research based articles were analyzed and 211 of these reported at least one magnitude of effect measure, a hit-rate of only 22.8%. Furthermore, it was commonly observed that, even when reported, these effect sizes were seldom used in interpretation of results. This finding paints a rather grim picture regarding effect size use as an alternative to SSTs. Finally, examples of other recommended reporting practices, such as internal replicability analyses (Daniel, 1998; Thompson, 1996), indexing statistically significant results to sample size (Thompson, 1994b; Vacha-Haase & Nilsson, 1998), and use of confidence

intervals (Kirk, 1996; McLean & Ernest, 1998) were almost nonexistent.

Four Suggestions for Improved Research Practice

Multiple authors have called for reforms that would facilitate improved research practice and decrease overreliance (or at least misuse) on SSTs (see e.g., Carver, 1978, 1993; Cohen, 1990, 1994; Daniel, 1998; Kirk, 1996; Thompson, 1996). Some have argued for an absolute abolishment of SSTs (Carver, 1978; Rozeboom, 1997; Schmidt, 1996; Schmidt & Hunter, 1997). Short of this extreme, others have recommended inclusion of additional research information to facilitate interpretation and correct use of SSTs (Daniel, 1998; Thompson, 1996). More conservatively, even supporters of SSTs often recognize that SSTs tend to be misused. For example, Levin (1998b), a SST advocate, noted:

> . . . statistical hypothesis testing, as is generally practiced, is not without sin. I too oppose mindless . . . manifestations of it. Such manifestations surely portray the practice of hypothesis testing at its worst. More forethought and restraint on the part of researchers would likely help to deflect much of the criticism concerning its misapplication. (p. 48)

Among the many suggestions, the four proposals noted here tend to be the most frequent. Two recommendations involve more precise use of SSTs and two involve inclusion of alternative information beyond SSTs.

Report AND Interpret Effect Sizes

First, and perhaps foremost, effect size indices should be both reported and interpreted in addition to SSTs (Keselman et al., 1998; Thompson, 1996). These indices allow researchers to examine the "practical" significance of their results (Kirk, 1996) and interpret the magnitude of differences between groups or relationships between variables. Effect sizes come in many forms (see e.g., Kirk, 1996; Snyder & Lawson, 1993), including corrected measures that attenuate the effect statistic by correcting for sampling error (i.e., sample size, number of variables used, and theoretical magnitude of effect in the population). Uncorrected measures (e.g., R^2 and η^2) reflect the maximized relationship between variables resulting from statistical analyses that capitalize on the unique variance present in a sample. Depending on the degree of sampling error present, these measures tend to overestimate the magnitude of effect.

Some are wary of effect sizes, since interpretation of effect sizes becomes in inherently subjective process that invokes both the values of the researcher using the measures and those of the social community. A sufficiently large effect in one study may not "make the cut" in another context. For example, high stakes research studies, perhaps in medical settings where lives are at stake, may tolerate smaller magnitude of effects given the potential positive outcome. Levin (1998b, p. 45) warned against the potential "bias" in effect size interpretation. On the other hand, Kirk (1996) suggested that when interpreting confidence intervals (another suggested reform),

> an element of subjectivity is introduced into the decision process. . . And the judgment [whether the result is trivial, useful, or important] inevitably involves a variety of considerations, including the researcher's value system, societal concerns, costs and benefits, and so on. (p. 755)

Of course, to assume that science is entirely objective is to exhibit psychological denial. In any research study, myriad decisions are made (by fallible humans) that potentially impact outcomes. The key here is to recognize this dynamic and proceed

with reflective, thoughtful judgment, which includes the defensible interpretation of effect sizes. Vacha-Haase and Nilsson (1998) noted correctly, "The issues of practical significance, generalizability, and replicability of results must always be interpreted with care" (p. 56). Such judgments are no more hurtful to a scientific enterprise than the pseudo-objectivity of SSTs (Thompson, 1999). As noted by the APA TFSI report (Wilkinson & TFSI, 1999): "Good theories and intelligent interpretation advance a discipline more than rigid methodological orthodoxy. . . . Statistical methods should guide and discipline our thinking but should not determine it" (p. 604).

Index Statistically Significant Results to Sample Size

Since the null hypothesis is always false (Cohen, 1994), a statistically significant result is possible at some sample size, regardless of the effect observed. As such, results of SSTs are more readily interpretable if they are explicitly indexed to the sample size used. Post hoc power analyses can shed light on (a) whether a researcher failed to reject the null simply because she lacked two more subjects or (b) whether it took 1,200 subjects to "find" a minute (unimportant?) difference between groups. Of course, power analyses are best considered

as an a priori process (Cohen, 1988) to help avoid the first of these problems.

Thompson (1989) suggested conducting so-called "what-if" analyses to interpret results in sample size context. For an observed sample effect in a study (e.g., R^2 and η^2), these analyses examine either at what sample size would a non-statistically significant result become so or at what sample size would a statistically significant result cease to exist. This information would help readers with SST interpretation.

These original "what-if" methods contained inherent weaknesses, however, because they are based on uncorrected effect size estimates. Kieffer and Thompson (1999) recently proposed a more precise "what-if" method that utilizes corrected effect sizes in the calculation of hypothesized p-values. Since sampling error varies based on the elements listed above (including sample size), the estimated population effect will also vary as sample sizes changes. As more subjects are included in a sample, the more accurate the uncorrected effect size tends to be. Less correction is then necessary to adjust for the population effect. Since the effect size directly impacts power, Kieffer and Thompson developed a method for using the corrected effect estimate in "what-if" analyses, which

facilitates more accurate interpretation of both statistical significance and magnitude of effect. Researchers are encouraged to utilize this newer method.

Provide Evidence of Replicability

By and large, science is about discovering theory that holds true, to some degree at least, in multiple applications. While some researchers from a more quantitative penchant may disagree with this purpose, quantitative methods concern themselves with unveiling results that are generalizable to populations of interest. No thoughtful researcher wants to announce his or her "groundbreaking" discovery too loudly until some evidence that the finding was not a fluke emerges. Unfortunately, SSTs have historically been misinterpreted as providing such evidence (Daniel, 1998; Shaver, 1993).

Put simply, there is no substitute for external replication with independent samples. However, in a literature biased against non-statistically significant results and less than novel investigations, replications are seldom reported. Of course, researchers themselves often tire of data collection and fail to replicate.

Fortunately, there are other ways to examine replicability, including comparing one's own obtained effect size to those

published in prior investigations. As suggested by Thompson and Snyder (1997),

> . . . explicitly and reflectively linking research results in a given study to the effect sizes in previous studies is also a vehicle for evaluating result replicability. This can be done prospectively by formulating null hypotheses incorporating specific parameter expectations derived from previous research, as against the contemporary practice of always testing hypotheses of no difference or of no relationship (i.e., what Cohen, 1994, described as "nil" hypothesis testing). (p. 80)

The APA TFSI report (Wilkinson & TFSI, 1999) concurred, "We must stress again that reporting and interpreting effect sizes in the context of previously reported effects is essential to good research. It enables readers to evaluate the stability of results across samples, designs, and analyses" (p. 599).

Internal replications can also be conducted. Most common of these analyses are the bootstrap, jackknife, and crossvalidation. All internal replications manipulate one's data in various ways and examine stability of statistics under variant sample conditions. Of course, these internal replications are ultimately based on the same data as the

sample, making replication assertions tentative. Knapp (1998) argued that these analyses only allow for "estimating sampling error without making the traditional parametric assumptions" (p. 39). Similarly, Levin (1998b) suggested that internal replications are

> nice for establishing the robustness of a single study's conclusions . . . However, that type of "replication" is neither as impressive nor as imperative for the accumulation of scientific knowledge as is a "replication" defined by an independently conducted study . . . (p. 47, emphasis in original)

While external replications are ideal, an estimation of sampling error in one's sample does speak to the potential generalizability of one's results. When sampling error is large, confidence in generalizability decreases, and vice versa. Short of conducting independent studies, providing evidence from internal replications is certainly superior than providing no evidence at all (Thompson, 1997).

Say "Statistically Significant," Not Just "Significant"

When an author states that her results were "significant" when she rejected the null hypothesis, too many persons equate such a statement to: "the results were important". Use of

precise language, by always stating "statistically significant", when referencing rejection of the null hypothesis may help guard against this misconception. While this recommendation is not as substantive as those preceding, appropriate language use will at least make clear when "the author intends to make claims about the "practical significance" (Kirk, 1996) of the results (Daniel, 1998, p. 29).

Conclusion

Null hypothesis significance tests have a storied history of abuse and misinterpretation (Huberty, 1993). Despite criticism for decades, many researchers still have misconceptions concerning what SSTs can and cannot do (Nelson, Rosenthal, Rosnow, 1986; Oakes, 1986; Rosenthal & Gaito, 1963; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993). Alternative means for interpretation, such as using magnitude of effect measures, have been recommended by many (e.g., Kirk, 1996) to assess practical significance. However, few researchers actually report effect sizes and even fewer interpret them. Important in this debate is the recent report from the APA TFSI (Wilkinson & TFSI, 1999) which stated that researchers should always report effect size measures when using p-values. Research practice and quality would benefit from always

reporting effect sizes, indexing statistically significant results to sample size, providing replicability evidence, and using precise language concerning a "statistically significant" result.

References

Abelson, R. P. (1997). A retrospective on the significance testing ban of 1999 (If there were no significance tests, they would have to be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 117-141). Mahwah, NJ: Erlbaum.

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington, DC: Author.

Azar, B. (1997). APA task force urges a harder look at data. APA Monitor, 28(3), 26.

Board of Scientific Affairs. (1996). Task Force on Statistical Inference initial report (DRAFT) [Online]. Available: http://www.apa.org/science/tfsi/html.

Burdenski, T. (1999, January). A review of the latest literature on whether statistical significance tests should be banned. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio. (ERIC Document Reproduction Service No. ED 427 084)

Burkson, J. (1942). Tests of significance considered as evidence. Journal of the American Statistical Association, 37, 325-335.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61, 287-292.

Cohen, J. (1988). Statistical power analysis (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Cohen, J. (1994). The earth is round ($p<.05$). American Psychologist, 49, 997-1003.

Cortina, J. M., & Dunlap, W. P. (1997). Logic and purpose of significance testing. Psychological Methods, 2, 161-172.

Daniel, L. G. (1997). Kerlinger's research myths: An overview with implications for educational researchers. Journal of Experimental Education, 65, 101-112.

Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. Research in the Schools, 5, 23-32.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. Psychological Methods, 1, 379-390.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. American Psychologist, 52, 15-24.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). What if there were no significance tests? Mahwah, NJ: Erlbaum.

Harris, M. J. (1991). Significance tests are not enough: The role of effect-size estimation in theory corroboration. Theory & Psychology, 1, 375-382.

Hays, W. L. (1981). Statistics (3rd ed.). New York: Holt, Rinehart and Winston.

Heldref Foundation. (1997). Guidelines for contributors. Journal of Experimental Education, 65, 95-96.

Huberty, C. J . (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. Journal of Experimental Education, 61, 317-333.

Keselman, H. J., Huberty, C. J, Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. Review of Educational Research, 68, 350-386.

Kieffer, K. M., & Thompson, B. (1999, November). Interpreting statistical significance test results: A proposed new "what if" method. Paper presented at the annual meeting of the Mid-South Educational Research Association, Point Clear, AL.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Knapp, T. R. (1998). Comments on the statistical significance testing articles. Research in the Schools, 5, 39-41.

Levin, J. R. (1993). Statistical significance testing from three perspectives. Journal of Experimental Education, 61, 378-382.

Levin, J. R. (1998a). To test or not to test H_o? Educational and Psychological Measurement, 58, 313-333.

Levin, J. R. (1998b). What if there were no more bickering about statistical significance tests? Research in the Schools, 5, 43-53.

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. Research in the Schools, 5, 15-22.

Measurement and Evaluation in Counseling and Development. (1992). Guidelines for authors. Measurement and Evaluation in Counseling and Development, 25, 143.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Murphy, K. R. (1997). Editorial. Journal of Applied Psychology, 82, 3-5.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. American Psychologist, 41, 1299-1301.

Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. Research in the Schools, 5, 3-14.

Nunnally, J. (1960). The place of statistics in psychology. Educational and Psychological Measurement, 20, 641-650.

Oakes, M. (1986). Statistical inference: A commentary for the social and behavioral sciences. New York: Wiley.

Pedhazur, E. J. (1997). Multiple regression in behavioral research: Explanation and prediction (3rd ed.). Forth Worth, TX: Harcourt Brace.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, 57, 416-428.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were no significance tests? (pp. 335-392). Mahwah, NJ: Erlbaum.

Rosenthal, R., & Gaito, J. (1963). The interpretation of level of significance by psychological researchers. Journal of Psychology, 55, 33-38.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. Psychological Methods, 1, 115-129.

Schmidt, F., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What if there were not significance tests? (pp. 37-64). Mahwah, NJ: Erlbaum.

Shaver, J. (1985). Chance and nonsense. Phi Delta Kappan, 67, 57-60.

Shaver, J. (1993). What statistical significance testing is, and what it is not. Journal of Experimental Education, 61, 293-316.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61, 334-349.

Snyder, P., & Thompson, B. (1998). Use of tests of statistical significance and other analytic choices in a school psychology journal: Review of practices and suggested alternatives. School Psychology Quarterly, 13, 335-348.

Thompson, B. (1989). Asking "what if" questions about significance tests. Measurement and Evaluation in Counseling and Development, 22, 66-68.

Thompson, B. (1993). Foreword. Journal of Experimental Education, 61, 285-286.

Thompson, B. (1994a). The concept of statistical significance testing. Measurement Update, 4, 5-6.

Thompson, B. (1994b). Guidelines for authors. Educational and Psychological Measurement, 54, 837-847.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25, 26-30.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. Educational Researcher, 26, 29-32.

Thompson, B. (1998a). In praise of brilliance: Where that praise really belongs. American Psychologist, 53, 799-800.

Thompson, B. (1998b). Statistical significance and effect size reporting: Portrait of a possible future. Research in the Schools, 5, 33-38.

Thompson, B. (1999). Statistical significance tests, effect size reporting, and the vain pursuit of pseudo-objectivity. Theory & Psychology, 9, 191-196.

Thompson, B., & Snyder, P. A. (1997). Statistical significance testing practices in The Journal of Experimental Education. Journal of Experimental Education, 66, 75-83.

Thompson, B., & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent JCD research articles. Journal of Counseling and Development, 76, 436-441.

Tukey, J. W. (1991). The philosophy of multiple comparisons. Statistical Science, 6, 100-116.

Tyler, R. W. (1931). What is statistical significance? Educational Research Bulletin, 10, 115-118, 142.

Vacha-Haase, T., & Nilsson, J. E. (1998). Statistical significance reporting: Current trends and uses in MECD. Measurement and Evaluation in Counseling and Development, 31, 46-57.

Vacha-Haase, T., & Thompson, B. (1998, August). APA editorial policies regarding statistical significance and effect size: Glacial fields move inexorably (but glacially). Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Wilkinson, L. & Task Force on Statistical Inference. (1999). Statistical Methods in psychology journals: Guidelines and explanation. American Psychologist, 54, 594-604.

Zuckerman, M., Hodgins, H. S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. Psychological Science, 4, 49-53.